BIGDATA Processing with Hadoop
(PIG Scripting)

Problem 1: NASA WebServer Log Activity Analysis

For this task you will use the logged activity from a NASA Space Center Web server.The log records are stored as an ASCII file with one line per request, with the format:

host timestamp request reply bytes

where,

- host is the address of the host making the request. The address may be a hostname when possible, or the IP address if the hostname could not be looked up.
- timestamp in the format "[DAY MON DD HH:MM:SS YYYY]", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The timezone is GMT-0400.
- request is escaped within quotes and contains the HTTP command, the file accessed, and the protocol being used.
- reply is the HTTP reply code.
- bytes is the number of bytes in the reply.

Given the above description the following line:
128.159.105.240 - - [03/Aug/1995:13:10:49 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 4673
is to be read as: host 128.159.105.240 made a request at 03/Aug/1995:13:10:49 -0400 to retrieve (GET) file /shuttle/countdown/ using protocol HTTP/1.0; the server replied with code 200 and the result was 4673 bytes long. Note that the IP address of the host is delimited with the rest of the record with '- -'; and the entire request is within quotes.

Write PIG scripts to answer the following questions:

a. Find the most popular page of the Web server. (The one that appears most frequently in requests, regardless of command.)

b. Find the top 10 hosts that produced the most 404 HTTP errors.

c. Find the time difference between the first and the last visit for each host (if a host has visited the server only once than just print the timestamp of the visit).

BIGDATA Processing with Hadoop (PIG Scripting)

Problem 2: Querying Stack Overflow posts

For this task you will use a dataset from StackOverflow and the dataset contains a number of post records. Each record is represented as an XML row element with a number of comma-delimited attributes key-value pairs, associated with the row. Each record has its own identifier stored in a field named Id and a type, indicated by the value of a field PostTypeId. If the value of PostTypeId is 1, than the post refers to a question, otherwise is the value of PostTypeId is 2 the post refers to an answer.

An example of a question post is:

<row Id="2155" PostTypeId="1" AcceptedAnswerId="2928" CreationDate="2008-08-05T12:13:40.640" Score="25" ViewCount="17551" Body="The question content" OwnerUserId="371" LastEditorUserId="2134" LastEditorDisplayName="stackoverflowGuy" LastEditDate="2008-08-23T18:09:09.777" LastActivityDate="2013-09-19T15:39:43.160" Title="How do I?" Tags="asp.net" AnswerCount="6" CommentCount="0" FavoriteCount="12" />

In this example, Id="2155" represents the unique identifier given to the post; PostTypeId="1" means that this post is a question; AcceptedAnswerId="2928" means that the accepted answer from the user for this query is the answer with Id="2928"; and so on.

An example of a post that corresponds to an answer is:
<row Id="659891" PostTypeId="2" ParentId="659089" CreationDate="2009-03-18T20:07:44.843" Score="1" Body="Description of the problem" OwnerUserId="45756" OwnerDisplayName="terminator" LastActivityDate="2009-03-18T20:07:44.843" CommentCount="0" />

The attribute-value pair Id="659891" represents the unique identifier given to this post. The value for ParentId represents the identifier of the question this answer applies to, and the value for OwnerUserId represents the user who wrote the answer for this question.

You do not need to know exactly what each attribute means, but you will need to be able to access the value of each attribute given a record. You will need to write PIG scripts to answer the following questions. For each question we give the expected output format; lines beginning with '%' are only descriptive comments and you do not need to print them; you only need to print the actual output values.

a. Which are the 10 most popular questions according to their view counts (attribute ViewCount in a question post)?
Output Format:
% Question Id, Count
659891, 17551
659892, 2131

b. Who was the user that answered the most questions and what were the Ids of these questions?
Output Format:
% OwnerUserId -> PostId, PostId, PostId, ...
1342 -> 23, 26, 531

c. Who was the user that had the most accepted answers and what were these answers?
Output Format:
% OwnerUserId -> Number of accepted answers, AnswerId, AnswerId, ...
1 -> 1, 1432, 1643

Problem 3: Working with YELP JSON DataSet

We will work on Yelp dataset that can be obtained from this link: http://www.yelp.com/dataset_challenge/. You can find the details of attributes under the section "Notes on DataSet" at above link. Test the following PIG script to convert json data into tsv format:

```
reviews = LOAD 'path to your input dataset' USING
JsonLoader('votes:map[], user_id:chararray, review_id:chararray,
stars:int, date:chararray, text:chararray, type:chararray,
business_id:chararray');

tabs = FOREACH reviews GENERATE (INT) votes#'funny', (INT)
votes#'useful', (INT) votes#'cool', user_id, review_id, stars,
REPLACE(REPLACE(text, 'n', ''), 't', ''), date, type, business_id;

STORE tabs INTO 'path to output directory';
```

Note: You can use JsonLoader function available natively or as part of elephant bird project.

BIGDATA Processing with Hadoop
(PIG Scripting)

Problem 4: "People you may know" Recommendation System

One of the most basic principles in a social network is the following:
"If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future." That is, if users B and C have a friend A in common, there is an increased likelihood that B and C will become friends themselves. This principle is called *triadic closure* because the edge formed by the friendship of B and C "closes" the triangle formed by the edges of the nodes A,B , and C. If you are user B in this example, user C is one of the "People You May Know".
Note that edges are undirected and, therefore, friendships are mutual: if A is friend with B, B is also friend with A. This is a common friendship system in social networks such as Facebook/LinkedIn.
The input file has one line of the following format for each user:

<user><TAB><comma-separated list of user's friends>

Write PIG script that recommends, for each user U, a list of top-10 people(who are not already friend with U) with whom U may connect with. The output file has one line of the following format for each user:

<user><TAB><comma-separated list of people the user may know>

You create a random sample dataset for this problem or you can find a sample data set here:
http://importantfish.com/wp-content/uploads/2013/07/soc-LiveJournal1Adj.txt